Visualisation et classification des parcours de vie

Nicolas S. Müller*, Sylvain Lespinats**, Gilbert Ritschard*, Matthias Studer*, Alexis Gabadinho***

*Département d'économétrie, Université de Genève
{nicolas.muller, gilbert.ritschard, matthias.studer}@metri.unige.ch
**INSERM Unité 722 et Université Denis Diderot
Paris 7, Faculté de médecine, site Xavier Bichat
sylvain.lespinats@bichat.inserm.fr
***Laboratoire de démographie, Université de Genève
alexis.gabadinho@ses.unige.ch

Résumé. Cet article propose une méthodologie pour la visualisation et la classification des parcours de vie. Plus spécifiquement, nous considérons les parcours de vie d'individus suisses nés durant la première moitié du XXème siècle en utilisant les données provenant de l'enquête biographique rétrospective menée en 2002 par le Panel suisse de ménages. Nous nous sommes concentrés sur ces événements du parcours de vie : le départ du foyer parental, la naissance du premier enfant, le premier mariage et le premier divorce. A partir des données de base sur ces événements, nous discutons de leur transformation en séquences d'états. Nous présentons ensuite notre méthodologie pour extraire de la connaissance des parcours de vie. Cette méthodologie repose sur des distances calculées par un algorithme d'optimal matching. Ces distances sont ensuite utilisées pour la classification des parcours de vie et leur visualisation à l'aide de techniques de « Multi Dimensional Scaling ». Cet article s'intéresse en particulier aux problématiques entourant l'application de ces méthodes aux données de parcours de vie.

1 Introduction

Nous proposons dans ce travail d'étudier et de comparer diverses techniques de visualisation et de classification de parcours de vie [1]. Plus spécifiquement, nous considérons les parcours de vie familiale d'individus suisses nés durant la première moitié du XXème siècle à partir de données récoltées par le Panel suisse de ménages. Les parcours de vie familiale sont composés d'événements constitutifs de la vie familiale, comme le départ du foyer parental, le premier enfant, le premier mariage ou le premier divorce. Il est possible, à partir de ces événements, de considérer des parcours de vie individuels sous la forme de séquences d'états, chaque événement survenant dans la vie de l'individu correspondant à un changement d'état.

[1] Etude soutenue par le Fonds national suisse de la recherche (FNS) FN-100012-113998, et réalisée avec les données collectées dans le cadre du projet « Vivre en Suisse 1999-2020 », piloté par le Panel suisse de ménages et supporté par le FNS, l'Office fédéral de la statistique et l'Université de Neuchâtel.

Une méthodologie ad hoc destinée à créer une typologie des parcours de vie et à visualiser les comportements individuels et l'évolution des normes sociales les régulant est présentée ici. La méthode principale consiste à calculer une distance entre chaque séquence à l'aide d'un algorithme d'optimal matching ; on obtient ainsi une distance qui respecte le caractère temporel des séquences de parcours de vie. Les résultats sont ensuite visualisés à l'aide de méthodes de type « Multi Dimensional Scaling ».

Cet article est construit de la manière suivante. La première partie présente les données utilisées ainsi que les transformations nécessaires pour construire des séquences d'états à partir d'événements. La deuxième partie présente la méthode d'optimal matching, son fonctionnement et la problématique de la définition du coût des opérations. La troisième partie concerne les méthodes de visualisation de type Multi Dimensional Scaling et leur principe de fonctionnement. La quatrième partie présente les résultats de l'application de notre méthodologie aux données du Panel suisse de ménages. Les résultats sont interprétés à l'aide de graphiques et de modèles de régression logistiques. Nous concluons finalement sur les possibilités que nous permettent d'envisager l'application de cette méthodologie aux données en sciences sociales.

2 Données

A partir des réponses à un questionnaire, nous extrayons des données sous la forme d'un tableau où chaque ligne est un individu et chaque colonne une variable (tableau 1).

TAB. 1 – *Exemple de données sous la forme d'événements*

ind.	naissance	départ	mariage	enfant	divorce
1	1974	1992	1994	1996	n/a

Le passage à une représentation sous forme de séquences d'états n'est pas trivial. La difficulté consiste à représenter sous la forme d'un unique état une combinaison d'événements qui se sont déjà produits ou non à chaque âge. De manière plus formelle, nous définissons l'état qui définit un individu à un âge précis comme une information sur les événements réalisés. On peut dire, à partir d'un état, quels événements se sont déjà produits. La réalisation d'un ou de plusieurs événement durant une année t entraîne le passage de l'état dans lequel se trouvait l'individu à $t-1$ à un nouvel état. La définition des états à partir des événements est un problème propre au type de données et à la problématique de recherche. Une manière simple de procéder consisterait à créer un état pour chaque combinaison d'événements. Avec cette solution, le nombre d'états s'élèverait à 2^n pour n événements, ce qui rend l'interprétation difficile dès lors qu'on prend en considération beaucoup d'événements. Nous avons donc choisi d'agglomérer certaines combinaisons en accord avec les objectifs de recherche.

Dans le cadre de cette étude, nous avons décidé de retenir quatre événements constitutifs de la vie familiale : le départ du foyer parental, le premier mariage, le premier divorce et la naissance du premier enfant. Le tableau 2 présente le codage des états que nous avons établi par rapport aux quatre événements retenus. Le nombre d'états a été réduit de 16 à 8, notamment en supprimant des états impossibles (tous ceux qui contiennent un divorce sans un mariage préalable) ou en combinant deux états (par exemple l'état 2 concerne les individus mariés qui

ne sont pas partis du foyer parental, qu'ils aient eu des enfants ou non). En se référant à cette liste d'états et à l'exemple donné dans le tableau 1, le résultat de la création d'une séquence de parcours de vie familiale se trouve dans le tableau 3.

TAB. 2 – *Liste des états*

	départ	mariage	enfant	divorce
0	non	non	non	non
1	oui	non	non	non
2	non	oui	oui/non	non
3	oui	oui	non	non
4	non	non	oui	non
5	oui	non	oui	non
6	oui	oui	oui	non
7	oui/non	oui/non	oui/non	oui

TAB. 3 – *Exemple de données sous forme de séquence d'états*

individu	1974	...	1991	1992	1993	1994	1995	1996	1997	1998	...
1	0	...	0	1	1	3	3	6	6	6	...

Les données utilisées dans ce travail proviennent de l'enquête biographique rétrospective menée par le Panel suisse de ménages (www.swisspanel.ch) en 2002. Nous n'avons gardé que les individus âgés d'au moins 45 ans au moment de l'enquête, afin de n'avoir que des séquences complètes entre 15 et 45 ans. Ainsi, notre échantillon est composé de 2601 individus nés entre 1909 et 1957.

3 Optimal matching

La méthode d'analyse de séquences que nous utilisons dans ce travail est celle dite d'optimal matching. L'algorithme retenu est inspiré des méthodes d'alignement de séquences et de programmation dynamique utilisées en biologie moléculaire, notamment pour la comparaison de protéines ou de séquences d'ADN supposées homologues (Deonier et al., 2005; Needleman et Wunsch, 1970). Ce type de méthode a été conçu pour permettre la comparaison rapide de nombreuses séquences afin de trouver des correspondances parmi celles-ci. Les premiers algorithmes d'optimal matching sont apparus au début des années 70 et leur première utilisation dans les sciences sociales remonte à l'article d'Abbott et Forrest sur leur application à des données historiques (Abbott et Forrest, 1986). On doit à Abbott de nombreux articles méthodologiques sur l'utilisation de ces méthodes dans les sciences sociales, et notamment en sociologie (Abbott et Hrycak, 1990; Abbott et Tsay, 2000). L'intérêt de l'application de cette méthode aux parcours de vie est de pouvoir ensuite procéder à une classification non supervisée en utilisant les distances calculées par l'optimal matching.

3.1 Méthode

Nous reprenons ici la formulation de Rohwer et Pötter (2002). Prenons Ω, l'ensemble des opérations possibles, et $a[w]$ le résultat de l'application des opérations $w \in \Omega$ sur la séquence a. Nous considérons trois types d'opérations : l'insertion d'un élément, la suppression d'un élément, ou la substitution d'un élément par un autre. Si l'on attribue un coût $c(w)$ qui correspond au coût d'appliquer l'opération $w \in \Omega$, la distance entre une séquence a et une séquence b peut être formalisée de la manière suivante : $d(a, b) = \min\{c[w_1, ..., w_k] \mid b = a[w_1, ..., w_k], w \in \Omega, k \geq 0\}$, avec $c[w_1, ...w_k] = \sum_{i=1}^{k} c[w_i]$. Autrement dit, pour chaque paire de séquences, on cherche la combinaison d'opérations pour rendre les séquences identiques dont la somme des coûts est la plus petite. L'algorithme utilisé pour trouver cette distance minimale utilise une méthode de programmation dynamique qui est décrite dans (Deonier et al., 2005). L'implémentation de l'algorithme que nous avons utilisée est celle présente dans le logiciel TDA ; son fonctionnement est détaillé dans son manuel d'utilisation (Rohwer et Pötter, 2002).

3.2 Définition des coûts

Comme nous l'avons vu précédemment, un coût c peut être attribué aux opérations $w \in \Omega$. Les coûts de substitution, auxquels nous nous sommes intéressés en particulier, peuvent être représentés sous la forme d'une matrice symétrique qui définit une valeur pour chaque paire d'état. L'attribution de ces valeurs en se basant sur un modèle théorique est particulièrement difficile dans le cadre d'une utilisation en sciences sociales, ce qui fait l'objet d'un débat (Wu, 2000). Il est en effet délicat de décider du coût du passage d'un état à un autre, mais il est pourtant intéressant et parfois capital de pouvoir différencier ces coûts. Pour cela, deux méthodes disponibles ont été essayées sur notre jeu de données. La première est implémentée dans le logiciel TDA (Rohwer et Pötter, 2002) et définit le coût de chaque substitution en fonction des taux de transition observés dans les données. Le coût du passage d'un état i à un état j est donc calculé de la manière suivante : $c_{i,j} = c_{j,i} = 2 - P(i_t|j_{t-1}) - P(j_t|i_{t-1})$. Le coût de base est fixé à 2, et plus la probabilité $P(i_t|j_{t-1})$ de passer de l'état i à l'état j, et inversement, est grande, plus ce coût baisse. Ainsi, les substitutions correspondantes aux transitions observées fréquemment seront moins coûteuse que celles qui n'arrivent jamais. Une autre méthode, proposée dans le logiciel T-COFFEE/SALTT (Notredame et al., 2005), consiste à calculer une matrice des coûts de substitution optimale par un processus itératif (Gauthier et al., 2007). Les tableaux 4 et 5 contiennent les résultats de l'application de ces deux méthodes de définition des coûts de substitutions sur nos données. Une analyse visuelle du tableau 4 permet d'observer qu'un passage de l'état 0 (aucun événement) à l'état 7 (divorce) ne s'observe jamais dans nos données, puisque son coût est de 2 dans les coûts tirés des taux de substitution (en gras). Cette transition correspondrait à un individu qui dans l'espace d'une année se marie puis divorce. Le passage de l'état 3 (départ et mariage) à l'état 6 (départ, mariage et enfant) est quant à lui beaucoup plus fréquent, et par conséquent moins coûteux. Le tableau 5 semble cohérent avec les coûts définis, même si la comparaison est difficile en raison de la plus grande variabilité des valeurs.

Le coût des opérations d'insertion et de suppression a quant à lui été fixé à une valeur unique de 3 dans la solution basée sur les taux de transition . Ce choix a pour but de favoriser au maximum les opérations de substitution (qui ont un coup maximum de 2) afin d'éviter les phénomènes de distorsion du temps qu'engendrent les opérations d'insertion. Avec cette

TAB. 4 – *Coûts de substitution (taux de transition)*

états	0	1	2	3	4	5	6	7
0	0	1.948	1.985	1.969	1.999	1.999	1.989	2
1	1.948	0	2	1.921	2	1.995	1.981	1.999
2	1.985	2	0	1.997	1.947	2	1.996	1.992
3	1.969	1.921	1.997	0	2	2	1.888	1.988
4	1.999	2	1.947	2	0	1.96	1.987	2
5	1.999	1.995	2	2	1.96	0	1.948	1.994
6	1.989	1.98	1.996	**1.888**	1.987	1.948	0	1.994
7	**2**	1.999	1.992	1.988	2	1.994	1.994	0

TAB. 5 – *Coûts de substitution (SALTT)*

états	0	1	2	3	4	5	6	7
0	0	0.881	19.253	1.785	14.913	17.045	20.769	18.679
1	0.881	0	1.702	0.774	1.519	1.185	1.3	2.15
2	19.253	1.702	0	1.25	0.8	1.313	1.34	1.351
3	1.785	0.774	1.25	0	1.083	0.988	0.855	1.189
4	14.913	1.519	0.8	1.083	0	0.901	1.232	1.715
5	17.045	1.185	1.313	0.988	0.901	0	1.064	1.319
6	20.769	1.3	1.34	**0.855**	1.232	1.064	0	0.936
7	**18.679**	2.15	1.351	1.189	1.715	1.319	0.936	0

solution, les seules situations où sont utilisées les insertions/suppressions sont en cas de léger décalage (p.ex. 0-1-2-3-4-4 à aligner avec 0-0-1-2-3-4). Dans le cas de la solution basée sur la matrice des coûts optimaux, le coût d'insertion/suppression a été fixé selon les recommandations de Gauthier et al. (2007), c'est-à-dire égal à la moyenne des coûts de substitution. La figure 1 donne une vision graphique de la disparité entre les matrices de distances calculées avec les différentes solutions de coût. Le graphique de gauche confronte les distances calculées avec les coûts de substitution fixés en fonction des taux de transition aux distances calculées avec un coût de substitution fixé à 2. Il apparaît très nettement que les résultats fournis par ces deux solutions sont quasiment identiques (fig. 1 partie gauche). La comparaison de la solution des taux de substitution avec la solution des coûts optimaux montre une plus grande disparité des distances et un effet d'échelle dû à la plus grande variabilité des coûts optimaux (fig. 1 partie droite). On peut en conclure qu'avec ce jeu de données, l'utilisation des taux de transition plutôt qu'un coût fixe n'a que peu d'influence sur les distances. En revanche, la différence entre la solution des taux de transition et la solution des coûts optimaux est plus marquée.

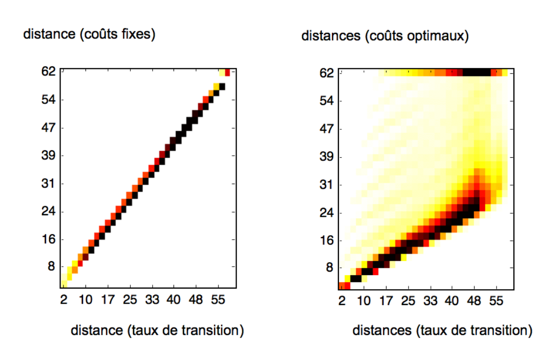


FIG. 1 – *La partie gauche présente les distances obtenues par la méthode avec les coûts substitutions basés sur les taux de transition selon les valeurs des distances obtenues avec des coûts de substitution fixés à 2 (le fait que les valeurs soient sur la diagonale indique que les distances obtenues par ces deux distances sont égales). La partie droite présente ces mêmes distances fondées sur les coûts de transition en fonction de celles calculées avec les coûts optimaux. Ces vues sont des graphiques en densité (plus la quantité de points associé à une unité de surface est grande, plus l'unité de surface est foncée), ainsi ces figures restent lisibles malgré la grande quantité de points présentés (environ* $2000^2/2$*).*

3.3 Classification

Nous sommes maintenant capables de produire une matrice de distances mesurant les différences entre les parcours de vie des individus. Celle-ci peut être utilisée dans une procédure de classification hiérarchique ascendante selon la méthode de Ward. Le tableau 6 croise les résultats obtenus par la classification hiérarchique ascendante entre les deux solutions à cinq groupes. Comme on peut le constater, la répartition des individus entre les groupes diffère fortement, même si certains groupes, comme le 4, semblent stable dans les deux solutions. Notre choix d'une solution à cinq groupes s'est faite en fonction de son interprétabilité, mais aussi à l'aide des méthodes graphiques présentées dans la partie suivante (Multi Dimensional Scaling). Nous avons choisi d'utiliser pour la suite de cet article les distances basées sur les taux de

TAB. 6 – *Croisement entre les deux solutions (taux de transition et coûts optimaux)*

taux de transition	coûts optimaux 1	2	3	4	5	Total
1	78	0	0	0	186	264
2	324	0	0	0	0	324
3	12	821	613	0	0	1446
4	37	1	7	259	0	304
5	1	253	3	0	6	263
Total	452	1075	623	259	192	2601

transition. Les résultats obtenus de cette manière sont plus facilement interprétables ; en effet, les groupes obtenus par le clustering sont plus homogènes et les coefficients des régressions logistiques plus significatifs.

4 Multi Dimensional Scaling

Avant de procéder à une classification hiérarchique ascendante, la matrice de distances apporte peu d'information aux experts. Ainsi, pour leur permettre d'appréhender les résultats, nous proposons de générer des « cartes » exprimant les relations de proximité entre les parcours des individus. Une telle représentation intuitive des données peut être obtenue par des méthodes de type « Multi Dimensional Scaling ». De cette manière, on dispose d'un outil qui permet de visualiser graphiquement les distances et d'aider à la décision du nombre de groupes à retenir dans une classification hiérarchique.

4.1 DD-HDS

Nous constatons que nos données présentent de fortes relations non-linéaires entre les individus. Ces relations mettent en échec les méthodes de projection linéaire (par exemple le Classical Multi Dimensional Scaling (Torgerson, 1952)). Pour parer à ce cas de figure, de nombreuses méthodes de représentation de données non-linéaires ont été proposées. Leur but commun est d'offrir une configuration de points sur un espace de faible dimension qui préserve les distances entre les données (avec un effort particulier pour la conservation des distances courtes). Parmi elles, nous avons choisi DD-HDS (Data-Driven High Dimensional Scaling, (Lespinats et al., 2007b)) pour sa capacité à éviter les « faux-voisinages » (données éloignées dans l'espace d'origine mais représentées comme proches) et les « déchirements » (données proches dans l'espace d'origine mais représentées comme éloignées).

Dans le cas des parcours de vie, nous sommes en mesure d'affirmer qu'une représentation en deux dimensions permet d'exprimer convenablement l'organisation des données. En effet lorsqu'elles sont projetées dans un espace tridimensionnel, on peut observer que les données occupent une surface bidimensionnelle (fig. 2). Les relations non-linéaires (matérialisées par le fait que cette surface ne ressemble pas du tout à un plan) sont ici manifestes.

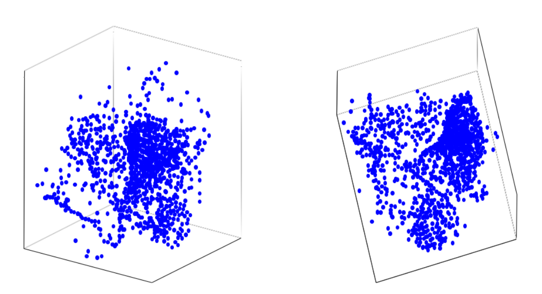


FIG. 2 – *Visualisation des données de parcours de vie dans un espace tridimensionnel (angles choisis). Chaque point correspond à un individu.*

Nous constatons en effet qu'une représentation sur un espace bidimensionnel permet de rapprocher les individus dont les parcours de vie sont proches. Par exemple on peut observer que les individus divorcés se rassemblent sur la droite de la représentation (fig. 3). Notons que plus le divorce est précoce, plus l'individu s'écarte vers la droite. La même analyse peut bien sûr être menée pour les 7 états, ce qui permet d'appréhender facilement l'organisation spatiale des individus (données non présentées).

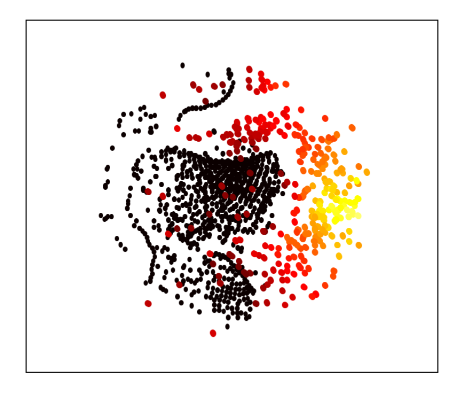


FIG. 3 – *Représentation bidimensionnelle des parcours de vie. Le code couleur permet de visualiser l'âge des divorces. Les points noirs de taille réduite correspondent aux individus qui n'ont pas divorcé. Le niveau de gris des autres points exprime l'âge de l'individu au moment du divorce. Plus l'individu est jeune au moment de son divorce, plus le point associé est clair.*

Ce type de représentation permet également de visualiser d'autres types d'informations. Par exemple, la figure 4 montre la répartition des dates de naissance des individus sur la représentation. Ainsi, nous observons que certains comportements ont eu tendance à disparaître comme le fait de rester chez ses parents (en haut au centre) et que des nouveaux comportements apparaissent comme les mariages tardifs (zone sur la gauche de la partie centrale).

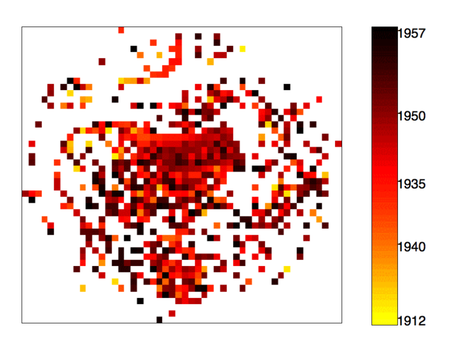


FIG. 4 – *Organisation des dates de naissance dans la représentation. La représentation est divisée en unité de surface, le niveau de gris de chaque zone dépend de la moyenne des dates de naissance (plus la date moyenne est ancienne, plus l'unité de surface associée est foncée).*

4.2 RankVisu

En termes de réduction de dimension, on cherche classiquement à préserver les distances entre données. RankVisu propose un nouveau point de vue sur les données en cherchant à conserver les rangs de voisinages (une relation de rang 1 est le relation entre un point et son plus proche voisin, une relation de rang 2 est la relation entre un point et son deuxième plus proche voisin, etc) (Lespinats et al., 2007a). Cette méthode renforce les groupes de données et permettra ainsi de valider notre clustering. La représentation obtenue à l'aide de RankVisu est mise en relation avec le résultat d'une classification hiérarchique (critère de Ward).

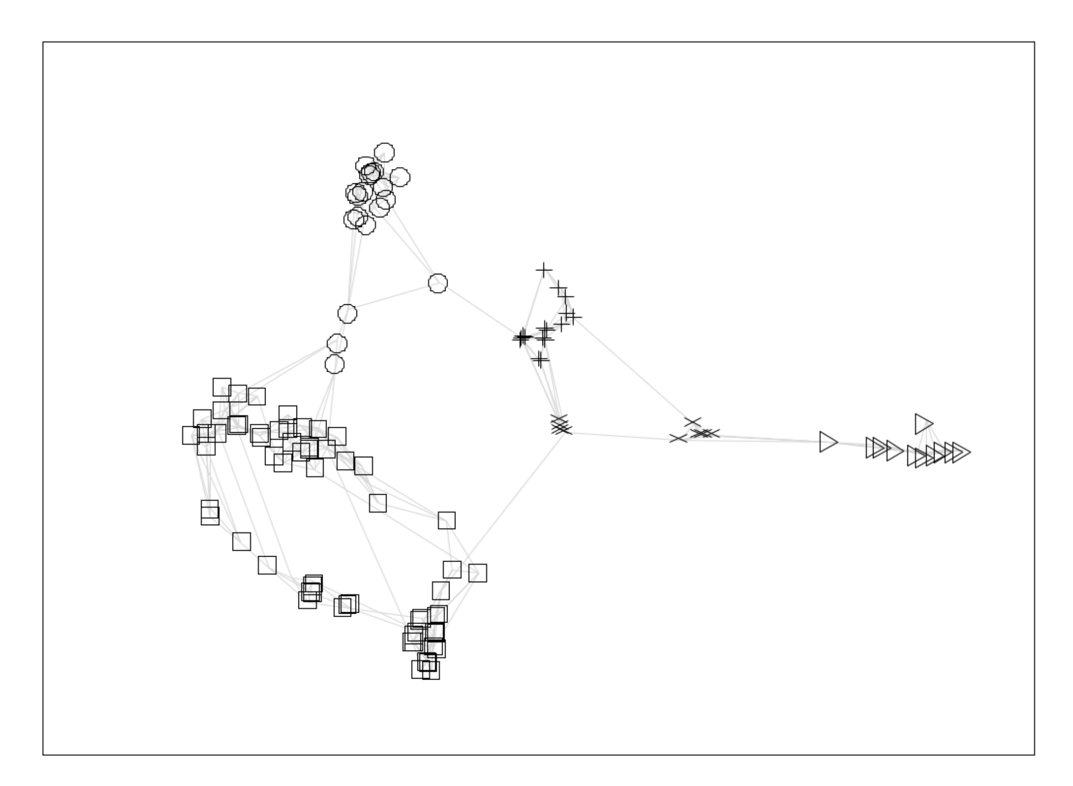


FIG. 5 – *Représentation bidimensionnelle (méthode RankVisu) de 100 individus représentatifs de la population (choisis par tirage au sort). Les cinq groupes d'appartenance sur la base de la classification hiérarchique sont exprimés par les signes. Les voisinages entre individus sont matérialisés par des segments qui relient chaque point à ces cinq plus proches voisins.*

Notons que ces deux méthodes se basent sur des informations relativement différentes : la classification s'appuie sur les distances tandis que RankVisu utilise les rangs de voisinage entre données. La figure 5 présente la représentation obtenue par RankVisu, en distinguant les groupes identifiés par la classification en cinq classes. Chaque classe forme sur le graphique un groupe bien défini, ce qui renforce le crédit de notre classification (fig. 5). En effet, les deux méthodes aboutissent à des conclusions comparables.

5 Interprétations

Nous analysons maintenant les caractéristiques de chacun des groupes. L'interprétation peut se faire de plusieurs manières ; nous privilégions ici une méthode visuelle pour la distinction des groupes. Nous disposons de deux types de graphique pour représenter la forme des séquences individuelles. Le premier type consiste à représenter, pour chaque âge entre 15 et 45 ans, la proportion d'individus se trouvant dans chaque état. La figure fig. 7 donne les représentations pour les groupes 2 et 4. Le deuxième type de graphique représente quant à lui chaque séquence individuelle. Ainsi, on lit sur l'abscisse l'âge de l'individu, et les séquences sont dessinées horizontalement. L'ordre dans lequel les séquences apparaissent est définie par la distance qui les sépare d'une séquence de référence choisie au hasard parmi toutes les séquences du groupe (fig. 8). Ce dernier type de graphique est réalisé à l'aide du module pour le logiciel Stata développé par Brzinsky-Fay et al. (2006).

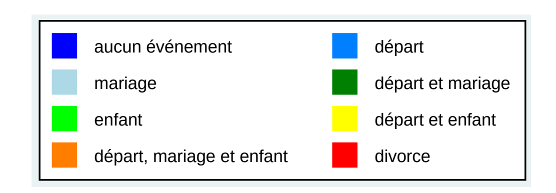


FIG. 6 – *Légende des couleurs des figures 7 et 8*

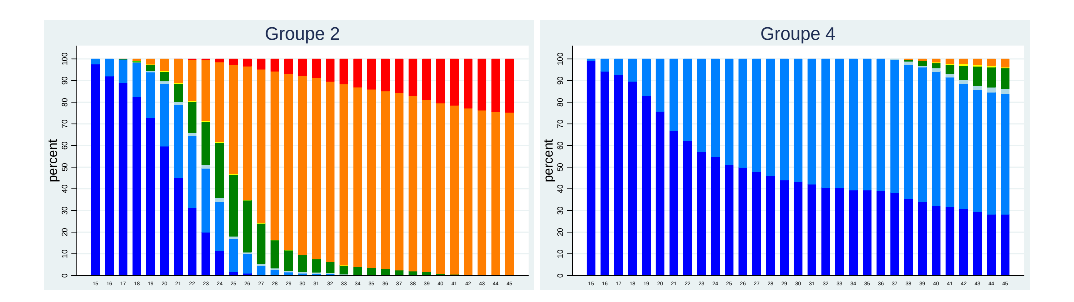


FIG. 7 – *Groupes 2 et 4 : proportions d'états*

D'un point de vue sociologique et démographique, les résultats de l'optimal matching permettent d'observer l'évolution dans le temps de certains phénomènes dégageant ainsi des effets de cohorte. Pourtant, dans le cas de cette application, il est difficile d'observer des changements de comportements étant donné la date de naissance maximale des individus. En effet, ceux-ci sont tous nés avant 1957, c'est-à-dire avant une période de modifications des comportements, c'est-à-dire la fin des années soixante. Les résultats bénéficieraient donc d'une réduction de la période de vie observée afin d'inclure des individus nés plus récemment. Les groupes sont malgré tout distinguable de manière graphique ; le groupe 4 (figure 7 correspond aux individus qui partent du foyer parental mais ne se marient pas ou tard (les premiers mariages interviennent à partir de 36 ans), alors que le groupe 2 contient des individus qui partent, se marient et ont des enfant jeunes (à 25 ils sont plus de 50% à en avoir).

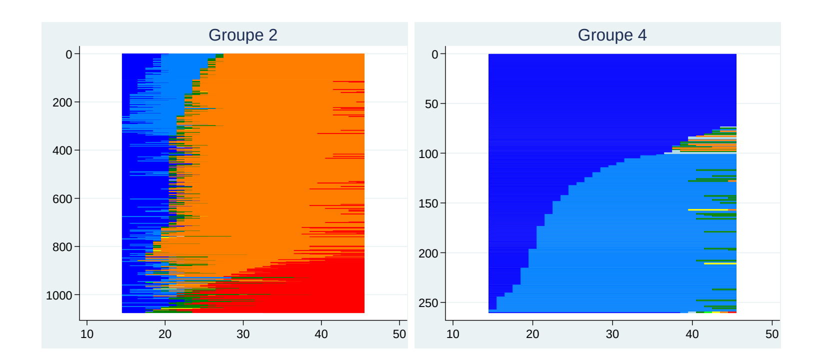


FIG. 8 – *Groupes 2 et 4 : séquences individuelles*

6 Conclusion

La méthodologie que nous mettons en place permet une fouille efficace des données de parcours de vie. Nous proposons une méthode performante pour quantifier les proximités entre parcours de vie, ainsi que des méthodes de visualisation qui permettent aux experts d'explorer les données de façon intuitive. Cette approche permet de prendre en compte les événements constitutifs d'un parcours de vie, qu'il soit professionnel, de santé, ou comme ici familial, et respecte leur ordre, leur durée, et par conséquent l'influence que ces événements peuvent avoir entre eux. L'assemblage des différents outils de visualisation donne à l'expert une connaissance des données qui serait autrement difficile à acquérir étant donné le nombre d'événements, d'années et d'individus pris en compte. Cette méthodologie met ainsi entre les mains des chercheurs qui possèdent des données longitudinales un puissant outil d'analyse exploratoire.

Références

Abbott, A. et J. Forrest (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History 16*, 471–494.

Abbott, A. et A. Hrycak (1990). Measuring resemblance in sequence data: An optimal matching analaysis of musician's carrers. *American Journal of Sociolgy 96*(1), 144–185.

Abbott, A. et A. Tsay (2000). Sequence analysis and optimal matching methods in sociology, Review and prospect. *Sociological Methods and Research 29*(1), 3–33. (With discussion, pp 34-76).

Brzinsky-Fay, C., U. Kohler, et M. Luniak (2006). Sequence analysis with stata. *The Stata Journal 6, number 4*, pp. 435–460.

Deonier, R., S. TavarŐ, et M. Waterman (2005). *Computational Genome Analysis: an Introduction*. Springer.

Gauthier, J.-A., E. D. Widmer, P. Bucher, et C. Notredame (2007). How much does it cost? Optimization of costs in sequence analysis of social science data. Manuscript, University of Lausanne. (Under review).

Lespinats, S., B. Fertil, P. Villemain, et J. Herault (2007a). Rankvisu : mapping from the neighbourhood network. submitted.

Lespinats, S., M. Verleysen, A. Giron, et B. Fertil (2007b). Dd-hds: a tool for visualization and exploration of highdimensional data. *IEEE transactions on Neural Networks Vol.18, 5*, pp. 1264–1279.

Needleman, S. B. et C. Wunsch (1970). General method applicable to the search for similarities in the animo acid sequence of two proteins. *Journal of Molecular Biology 48*, pp. 443–453.

Notredame, C., P. Bucher, J.-A. Gauthier, et E. Widmer (2005). T-COFFEE/SALTT: User guide and reference manual. disponible sur http://www.tcoffee.org/saltt.

Rohwer, G. et U. Pötter (2002). TDA user's manual. Software, Ruhr-Universität Bochum, Fakultät für Sozialwissenschaften, Bochum.

Torgerson, W. (1952). Multidimensional scaling: 1. theory and method. *Psychometrika vol. 17*, pp. 401–419.

Wu, L. (2000). Some comments on "sequence analysis and optimal matching methods in sociology : Review and prospect". *Sociological Methods and Research vol. 29(1)*, pp. 41–64.

Summary

This article proposes a methodology for visualizing and classifying life courses. More specifically, we consider life courses of Swiss people who lived during the 20th century, using data from a retrospective survey conducted in 2002 by the Swiss Household Panel. We focus on the following important events of the familial life: leaving parental home, having a child, getting married and divorcing. We first discuss how the original time stamped event data are transformed into an equivalent state sequence form. We present then our methodology for discovering useful knowledge from the observed life courses. It relies on distances based on an optimal matching algorithm. These distances are then used for clustering the life courses and visualizing them through Multi Dimensional Scaling techniques. The paper pays special attention to specific issues in the application of these methods to life course data.